Assessing geographic relevance for mobile search: A computational model and its validation via crowdsourcing

Reichenbacher, Tumasch ; De Sabbata, Stefano ; Purves, Ross S ; Fabrikant, Sara I

Abstract: The selection and retrieval of relevant information from the information universe on the web is becoming increasingly important in addressing information overload. It has also been recognized that geography is an important criterion of relevance, leading to the research area of geographic information retrieval. As users increasingly retrieve information in mobile situations, relevance is often related to geographic features in the real world as well as their representation in web documents. We present 2 methods for assessing geographic relevance (GR) of geographic entities in a mobile use context that include the 5 criteria topicality, spatiotemporal proximity, directionality, cluster, and colocation. To determine the effectiveness and validity of these methods, we evaluate them through a user study conducted on the Amazon Mechanical Turk crowdsourcing platform. An analysis of relevance ranks for geographic entities in 3 scenarios produced by two GR methods, 2 baseline methods, and human judgments collected in the experiment reveal that one of the GR methods produces similar ranks as human assessors.

DOI: https://doi.org/10.1002/asi.23625

Assessing geographic relevance for mobile search: a computational model and its validation via crowdsourcing

Tumasch Reichenbacher (corresponding author)

Department of Geography, University of Zurich, Winterthurerstr. 190, 8057 Zurich, Switzerland; tumasch.reichenbacher@geo.uzh.ch

Stefano De Sabbata

Oxford Internet Institute, University of Oxford, 1 St Giles, Oxford OX1 3JS, United Kingdom; stefano.desabbata@oii.ox.ac.uk

Ross S. Purves

Department of Geography, University of Zurich, Winterthurerstr. 190, 8057 Zurich, Switzerland; ross.purves@geo.uzh.ch

Sara I. Fabrikant

Department of Geography, University of Zurich, Winterthurerstr. 190, 8057 Zurich, Switzerland; sara.fabrikant@geo.uzh.ch

Abstract

The selection and retrieval of relevant information from the information universe on the web is becoming increasingly important in addressing information overload. It has also been recognised that geography is an important criterion of relevance, leading to the research area of geographic information retrieval. As users increasingly retrieve information in mobile situations relevance is often related to geographic features in the real world as well as their representation in web documents. We present two methods for assessing geographic relevance (GR) of geographic entities in a mobile use context that include the five criteria *topicality*, *spatio-temporal proximity*, *directionality*, *cluster*, and *co-location*. To determine the effectiveness and validity of these methods, we evaluate them through a user study conducted on the Amazon Mechanical Turk crowdsourcing platform. An analysis of relevance ranks for geographic entities in three scenarios produced by two GR methods, two baseline methods, and human judgements collected in the experiment reveal that one of the GR methods produces similar ranks as human assessors.

Keywords: geographic relevance, crowdsourcing, evaluation.

Assessing geographic relevance for mobile search: a computational model and its validation via crowdsourcing

Introduction

Geographic information is everywhere. The rise of mobile search, whose volume is now reported to have surpassed that of desktop search [1], makes the importance of differentiating between relevance in a mobile, location-based from a static, information seeking context all the more important (Raper, 2007; Reichenbacher, 2007). For many everyday activities related to, and often occurring in space (e.g. looking for a bar that is still open or finding an optimal route home in congested traffic) users require search results which take account of local context. Furthermore, many location-based services rank information, and display only a subset of possible options to users to display uncluttered results, often on a map, often simply using a combination of distance and feature type for ranking. Essentially, this task can be seen as analogous to that of traditional search – users require *geographically relevant* information about geographic entities in their environment in order to make appropriate choices. However, traditional information retrieval (IR) and geographic information retrieval (GIR) approaches which have been predominantly concerned with retrieving documents, fall short in addressing factors relevant to mobile search.

Taking account of local, geographic context, and retrieving *geographically relevant* information for users in a given context is thus a key challenge for information science. Mobile users often seek to solve spatial problems or answer spatial questions *in* the physical world and therefore establish relationships between spatial concepts in their mind, objects in physical space, and their representation on a mobile device. This degree of situatedness goes beyond topicality and includes concepts which are particular to information seeking in a

[1] http://searchenginewatch.com/sew/news/2411038/mobile-surpasses-desktop-in-search-queries

mobile context, such as personal mobility opportunities and limitations, environmental factors, simultaneously available activities, and affordances of places in the physical world. Having recognised this need, the notion of geographic relevance (GR) was introduced in Geographic Information Science (Raper, 2007). GR refers to the *relevance of a representation of a geographic entity* (i.e., a physical entity, such as a restaurant or a mountain), given a specific context of interaction with its representation, such as a point of interest on a digital map embedded in a specific, typically mobile, usage context. GR is thus expressed as the relation between a geographic entity and a human information need. Thus, although geographically referenced documents and documents containing geographic information may be a source of information in judging the GR of an entity, they are not the objective of the relevance assessment.

De Sabbata & Reichenbacher (2012) carried out a user study to identify five criterion which appeared to be of particular importance in calculating GR: *topicality*, *spatio-temporal proximity*, *directionality*, *cluster*, and *co-location*. Subsequently, De Sabbata (2013) developed a computational model to calculate GR based on these criteria. However, evaluating such a model requires, as is typical in information retrieval, some form of relevance judgements. Since, to our knowledge, no suitable benchmark data currently exist for mobile search, we chose to construct a set of scenarios for mobile search in three realistic usage scenarios. For each scenario we calculated GR using the five criterion listed above and baseline IR methods.

To judge the ranked lists, we opted to use a *crowdsourcing* approach. Crowdsourcing (Howe, 2006), is the outsourcing of, usually relatively simple, tasks to a large group of people. Participation is voluntary, and depending on the task and the interest of users, financially recompensed. Participants are assumed to work independently from one other, and cannot see the results of another's work. Crowdsourcing has, with certain limitations, already been used effectively in assessing relevance in IR (Alonso, Rose, & Stewart, 2008). However,

to our knowledge it has not been used to judge geographic relevance in the context of mobile search. Thus, in this paper we aim to address the following core research questions:

- Is a computational model of geographic relevance, including *topicality*, *spatio-temporal proximity*, *directionality*, *cluster*, and *co-location* more effective in ranking geographic entities than a model based on topicality and spatial proximity?
- Is crowdsourcing an appropriate approach to evaluating GR?
- Does a computational model of geographic relevance outperform baseline IR approaches to ranking geographic entities?

In the following we present firstly related work, before describing the methods by which we calculated GR and our crowdsourced relevance judgements. We then present and interpret our results, before discussing them in the context of our research questions and their broader implications.

Related work

Relevance criteria

Saracevic (1996) distinguishes five manifestations of relevance:

(1) the system or algorithmic relevance independent of the context and measures how well the query topic and document topic match;

(2) topical or subject relevance (aboutness, topicality);

(3) cognitive relevance or pertinence (informativeness, novelty);

(4) situational relevance or utility (usefulness in decision making, reduction of uncertainty); and

(5) the motivational or affective relevance (satisfaction, success).

A major distinction has to be made between objective (1) and subjective (2-5) relevance. The former has a long history of use in IR as a measure for the effectiveness of the retrieval process and is typically captured in precision and recall. The underlying assumption of the system or algorithmic relevance is that a system is capable of independently assessing the relevance of documents from the user, i.e. objectively. Many researchers (e.g., Cosijn & Ingwersen, 2000; Saracevic, 1996) suggested a more flexible approach going beyond simple binary relevance by measuring the semantic similarity of terms found in documents and query terms and then ranking the documents accordingly.

Geography has not played a major role in IR for long. One of the first in the field of IR to acknowledge a kind of spatial relevance was (Wilson, 1973) with the concept of situational relevance. However this concept was still targeted to classic document-based IR. More recently the interdisciplinary field of Geographic Information Retrieval (GIR) has studied the retrieval of documents where the query and documents retrieved contain spatial references and are often connected through spatial relations. The focus of GIR is however still on documents, and thus relevance in GIR is typically understood as relevance of documents with a spatial reference (Andrade & Silva, 2006; Cai, 2011; Clough, Joho, & Purves, 2006; Jones & Purves, 2008; Kumar, 2011).

The need for a comprehensive concept of GR applying to representations of real world geographic features was largely ignored until mobile technology matured enough and 3G mobile networks were implemented making Location Based Services (LBS) viable propositions (Raper, Gartner, Karimi, & Rizos, 2007b). Many LBS provide simple relevance filtering (Raper, Gartner, Karimi, & Rizos, 2007a) based on a user's current position (e.g., show the nearest 10 restaurants). However, such approaches have been criticised for their simplistic and narrow approach to relevance, which is effectively binary and based on spatial containment, e.g., by (Raper, 2007; Reichenbacher, 2007).

A concept of relevance for mobile information access rooted in IR was proposed by (Coppola, Della Mea, Di Gaspero, & Mizzaro, 2004). Although based on the notion of situational relevance (Saracevic, 1996, 2007; Wilson, 1973), and referring to the relevance of objects in the physical world with respect to a user's context, its main focus was not on geographic.

Several researchers have suggested that location alone is not sufficient for capturing a mobile user's context, and claim that there are other fundamental dimensions (Jiang & Yao, 2006; Raper et al., 2007b; Schmidt, Beigl, & Gellersen, 1999), such as time (Raubal, Miller, & Bridwell, 2004), activity (e.g., Crowley, Coutaz, Rey, & Reignier, 2002; Huang & Gartner, 2009), movement, and visibility (Mountain & Macfarlane, 2007). Other work has tried to go beyond LBS and its technological focus, and study mobile geographic information usage from a more fundamental perspective. Zipf (2003) was probably the first to introduce the idea of relevance of geographic entities on maps. He proposed a simple abstracted function for computing this relevance as a weighted linear combination of user characteristics and context parameters (e.g., spatial relevance, topical relevance etc.), although did not detail the parameters to be used. Raper (2007) discussed a very high-level and abstract perspective on GR. His conceptualisation of GR encompassed an individual 'attention' and an 'influence' dimension, as well as their relations. The individual attention is an expression of geographic information needs, which may be either allocentric or egocentric. The influence dimension of geographic objects in the environment is either space or place related.

The importance of the role of geography has been stressed (Raper, 2007; Reichenbacher, 2007) and spatial relationships have been proposed, such as, for example, spatial clusters for assessing the neighbourhood of a relevant entity in terms of other similar entities nearby , or co-location by analysing typical patterns of nearby entities belonging to different categories (Huang, Shekhar, & Xiong, 2004).

Combining relevance scores

Using the arithmetic product to combine individual relevance values (e.g., spatial relevance, topical relevance etc.) and compute GR does not seem likely to be a valid approach, since this method is non-compensatory, i.e., one low score is enough to yield a low aggregated score. This could generate invalid results, as the strong "and-ness" of the method would cause possibly relevant entities to be scored as absolutely irrelevant. To avoid the drawbacks of a simple arithmetic product in the field of GIR, Van Kreveld et al. (2005) and Purves et al. (2007) applied a geometric, compensatory, combination method, taking account of both thematic and geographic relevance. However, nonetheless it may still rank topically non-relevant documents highly, where no other documents are available by overestimating the importance of the geographic environment component.

A more flexible method that can deal with these drawbacks is the Continuous Preference Logic (CPL) model introduced by (Dujmovic, 1975, 2007), based on the generalised conjunction/disjunction (GCD) function (Dujmovic & Larsen, 2007). The core idea is the creation of logic operators with any grade of "and-ness" and "or-ness" in the range of [0, 1]. These are used by CPL to specify conjunctive partial (CPA) and disjunctive partial operators (DPA) for combining "mandatory" input with "desired" input in a conjunctive manner, and "sufficient" input with "desired" input in a disjunctive manner, respectively.

Evaluation of relevance and crowdsourcing

Evaluating GR brings with it several problems. Large-scale evaluations in IR often use benchmarks, i.e., test collections of documents for which relevance has been assessed by human experts. For GR, no such benchmarks are available. Although the Contextual Suggestion Track of TREC 2012 shares similar goals, it is not applicable to GR, as the adopted descriptions of context and granularity of spatio-temporal information are too coarse, and it was established for document-based IR. Similar problems apply to evaluation testbeds developed in GIR, such as GeoCLEF (Mandl, 2011). Moreover, large evaluation campaigns,

such as TREC, are not always affordable for small interest groups, focused on subfields of IR. In the last few years, crowdsourcing (Eickhoff & de Vries, 2011; Yuen, King, & Leung, 2011) has emerged as an alternative route to IR evaluation (Alonso & Baeza-Yates, 2011; Alonso & Mizzaro, 2009; Alonso, Rose, & Stewart, 2008; Carvalho, Lease, & Yilmaz, 2011). Crowdsourcing has been applied to particular IR tasks, such as video annotations (Soleymani & Larson, 2010), music similarity assessment (Urbano, Morato, Marrero, & Martín, 2010), and news search (McCreadie, Macdonald, & Ounis, 2010). More recently crowdsourcing has been applied in geographic information science for evaluating spatial formalisms of simple spatial overlap situations (Wallgrün, Yang, & Klippel, 2014). Such evaluations can be effectively crowdsourced through commercial providers such as Amazon Mechanical Turk (AMT), where any questionnaire or user experiment, that can be incorporated into a Web page, can be run as an Internet service without the need for further equipment. These platforms provide evaluators with tools to create and submit small *Human Intelligence Tasks* (HITs in AMT) to a wide audience of registered users, known as workers in AMT. As the tasks are rather often short and simple, and the number of workers is large, response times are commonly rather short.

While crowdsourcing platforms allow collection of human judgments for small datasets with a large sample, short response time, and at low costs, crowdsourced evaluations also have drawbacks. Little to nothing is known about the workers, although they can be assumed to be competent computer users or at least familiar with the Web environment. Studies performed on the demographics of AMT (Ipeirotis, 2010a, 2010b; Ross, Irani, Silberman, Zaldivar, & Tomlinson, 2010) showed that in 2008 most of the workers were residents of the Unites States of America, whereas participants from India accounted for almost half of the population in early 2010. The same study reported that about two thirds of the participants from India have at least a Bachelor's degree, and about one third declared that the money gained through AMT is at least sometimes 'necessary to make basic ends meet'.

This raises ethical concerns, as discussed by (Felstiner, 2011). At the same time, AMT offers an unique opportunity to perform experiments with such a diverse set of subjects (Mason & Suri, 2012). The control of answer quality is delicate (Marsden, 2009), since there is no guarantee that participants will carry out the tasks in a reliable and foreseen manner. However, malevolent workers can be discouraged by asking open questions or by giving more complex tasks (Eickhoff & de Vries, 2011; Harris, 2011). Today, most crowdsourcing platforms check the quality by evaluating the HIT approval rating of participants from previous experimenters. Short training phases prior to complex tasks and warm-up questions could further help to avoid potential misunderstandings, reduce errors and have shown to improve the answer quality. Clough, Sanderson, Jiayu, Gollins & Warner (2013) found limits of crowdsourcing in the evaluation of domain-specific search. Crowdsourcing workers and experts rank results very similarly, but workers seem, perhaps unsurprisingly, less able to differentiate levels of highly accurate search results.

Methods

Models for assessing the geographic relevance of geographic entities

De Sabbata (2013) developed a computational model for calculating geographic relevance (Figure 1) based on five relevance criteria *topicality*, *spatio-temporal proximity*, *directionality*, *cluster*, and *co-location*. The feasibility of the selection of these five criteria for computing a geographic relevance score was explored in previous studies by the authors (De Sabbata & Reichenbacher, 2012).

To compute the geographic relevance score of a geographic entity, we first operationalise the elicited criteria by mapping each criterion to a distance function. These distance functions (δ) take for each criterion a user query and a geographic entity as input. The computed distance increases as the relevance of a geographic entity in a specific user context diminishes (see Figure 1, left side). Furthermore, we normalise on an interval scale between 0 (no relevance) and 1 (maximal relevance) to obtain quantitative relevance scores.

Below we give a concrete example for the criterion of spatio-temporal showing how we compute a distance and the respective relevance score. We assume that a user submitting a space-time query at a given location in space and time, wishes to reach a destination within a given time. Based on the concept of space-time prisms (Miller & Bridwell, 2009), we take into account the time needed to reach an entity, and a time budget available to perform an activity at that location, also considering the temporal availability of the entity. We thus define the δ_{STprox} distance function as a ratio between the time needed to fulfil an activity, and the time a user is able to spend at a location of a given entity, while the entity is also available in terms of time of day (Equation 1):

$$\delta_{\text{STprox}}(q, g) = \frac{time\ needed}{time\ available} \textbf{ (Equation 1)}$$

As we also assume that utility grows less than linearly with decreasing distance, we define an auxiliary function d_{STprox} as a square root function of the inverse of the distance (see Equation 2). Utility is zero if an entity is not available for the time a user has available to accomplish the activity.

$$d_{\text{STprox}}(q, g) = \begin{cases} 0, & if\ \delta_{\text{STprox}} > 1 \\ \sqrt{\frac{1}{\delta_{\text{STprox}}}}, & otherwise \end{cases} \textbf{ (Equation 2)}$$

In order to obtain normalised scores for the distance values for spatio-temporal proximity, d_{STprox}, we define following normalisation function (Equation 3):

$$\bar{s}_{\text{STprox}}(q, g) = \frac{d_{\text{STprox}}(q,g)}{max_{j \in G}(d_{\text{STprox}}(q,j))} \textbf{ (Equation 3)}$$

The computation of the distance functions for the criteria *topicality*, *directionality*, *cluster*, and *co-location* follows a similar approach (De Sabbata, 2013).

Next, we compute a *mobility* score, by combining spatio-temporal proximity and directionality scores. For this we use Continuous Preference Logic (CPL) (Dujmovic, 2007) (see Equation 4).

$$\bar{s}_{Mobility}(q, g) = CPA_{0.75\ 0.75}\ (\bar{s}_{\text{STprox}}(q, g), \bar{s}_{\text{Direct}}(q, g)) \textbf{ (Equation 4)}$$

The conjunctive partial absorption (CPA) operator combines the "mandatory" spatio-temporal score with the "desired" directionality score in a conjunctive manner, i.e. the spatio-temporal score is a starting value and incremented or decremented, depending on whether the directionality score is greater or lower than the spatio-temporal score, and on the "and-ness" of the partial conjunction. If spatio-temporal score is zero, the combined output will also be zero. The scores for cluster and co-location are to a score for the geographic environment, respectively.

$$\bar{s}_{GeoEnv}(q, g) = \bar{s}_{\text{Clust}}(q, g) \; \Delta_{0.75} \; \bar{s}_{\text{Coloc}}(q, g) \quad \textbf{(Equation 5)}$$

Finally, these two aggregated scores *mobility* and *geographic environment* are combined with the topicality score to give a geographic relevance score (GR, Fig. 1), in the following referred to as ScoreGR (Table 1).

$$\text{GR}(q, g) = CPA_{0.75\ 0.75}\left(\bar{s}_{\text{Topicality}}(q, g), \bar{s}_{\text{Mobilityt}}(q, g), \bar{s}_{\text{GeoEnv}}(q, g)\right) \quad \textbf{(Equation 6)}$$

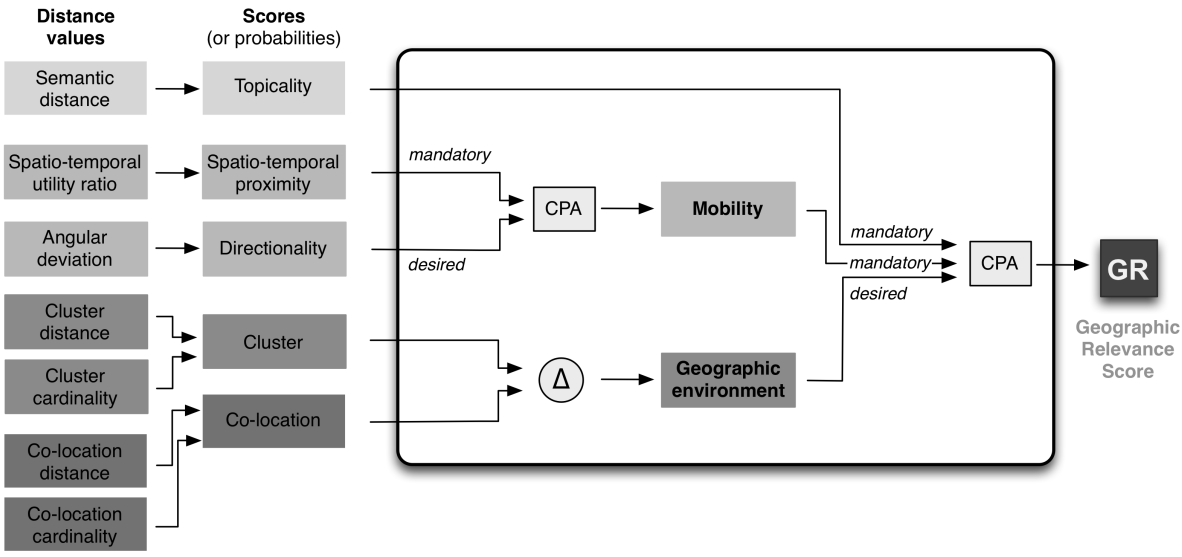


Figure 1: Computational model of geographic relevance

We also developed an alternative assessment method for GR that takes into account the distribution of the distance values of the geographic entities (De Sabbata, 2013). For this purpose, instead of scores, we calculate probabilities for the distribution of the distances for

the same five criteria as for *ScoreGR*. In analogy to the Okapi BM25 model (Spärck Jones, Walker, & Robertson, 2000) we refer to this method as *GRBM25* (Table 1). The similarity function is defined as follows:

$$sim_i(c,g) = \log\left[\frac{\|G\|}{odf(\delta_i,c,g)}\right] \cdot \frac{(k_1+1)\cdot d_i(c,g)}{k_1\cdot\left((1-b)+b\cdot\left(\frac{\delta_i(c,g)}{avg(\delta_i,c,G)}\right)\right)+d_i(c,g)} \quad \textbf{(Equation 7)}$$

where c ∈ C be a user context description, G = {g_1,g_2,...} a set of geographic objects, δ_i is a distance function and d_i is the related inversely proportional function. k_1 and b are tuning parameters derived from the original Okapi BM25 formula, and

$$\text{avg}(\delta_i, c, G) = \left[\frac{1}{\|G\|}\right]\sum_{g\in G}\delta_i(c,g) \quad \textbf{(Equation 8)}$$

$$\text{odf}(\delta_i, c, g) = \|\{h \in G | \delta_i(c,h) \leq \delta_i(c,g)\}\| \quad \textbf{(Equation 9)}$$

The first auxiliary function computes the average distance for a given context dimension c, while the latter computes the number of objects with equal or shorter distance to a user for a given object.

For comparison, we include two additional methods as baselines relying on simpler assessment models. The first baseline method reflects a simple LBS approach and will be referred to as *Baseline1* (Table 1) in the following. Given a query, *Baseline1* filters out all entities whose category does not match a user query, and orders the remaining entities according to the length of user's movement path (i.e., the distance from the user's current location to the location of the entity, and then to the destination). The second baseline method is referred to as *Baseline2* (Table 1). *Baseline2* combines the *topicality* score with a distance score computed as the inverse of the length of a user's path (i.e., normalised in the range [0,1], dividing it by the maximum obtained value), using the geometric combination method employed in the SPIRIT Project (Purves et al., 2007). Table 1 summarises the methods of GR assessment tested in the experiment and the relevance criteria included in our proposed methods.

Method	Criteria	Score
Baseline 1	topicality spatial proximity	category-based filter order by user's path length
Baseline 2	topicality spatial proximity	geometric combination (Purves et al., 2007) normalised inverse value of the user's path length
ScoreGR	topicality spatio-temporal proximity directionality cluster co-location	see (De Sabbata, 2013)
GRBM25	topicality spatio-temporal proximity directionality cluster co-location	see (De Sabbata, 2013)

Table 1: GR assessment methods tested in the experiment

Our evaluation follows the common benchmark-based testing of IR systems, where system relevance output is compared to relevance from judgements from humans, set as a benchmark. However, as no applicable benchmark to test the effectiveness of our proposed GR methods exists, we use crowdsourcing as an alternative evaluation method and employ the adopted approach by (Urbano et al., 2010) which uses simple pair-wise comparisons of preference judgements of similar music pieces against an item chosen as pivot and also allows for judging items as equally relevant. In order to evaluate the validity and effectiveness of these two GR assessment methods, *ScoreGR* and *GRBM25*, we designed a user-based experiment. We defined effectiveness as the similarity between the relevance rank produced by our GR computational model and the entities ranked using relevance judgements

performed by experiment participants through crowdsourcing which we consider as ground truth.

Experiment

A first objective of the experiment is to validate our proposed *ScoreGR* and *GRBM25* methods against human relevance assessments. A second objective is to establish whether the baseline methods provide a sufficient approximation of GR, even if they do not explicitly implement the criteria *spatio-temporal proximity*, *cluster*, and *co-location* as in *ScoreGR* and *GRBM25*. To meet these objectives, we designed three scenarios of mobile information seeking, involving (1) clusters of geographic entities (supermarkets, hotels, restaurants), (2) co-location rules (pharmacies next to supermarkets), and (3) spatio-temporally inaccessible entities (supermarkets, restaurants). We chose these three scenarios to balance the complexity of our evaluation tasks between tasks that reflect the core of GR (i.e., cluster, directionality, co-location), tasks that are atomic enough to be understood, tasks that are doable by the assumed population of workers, and tasks that still are ecologically valid. We did not consider simpler scenarios (e.g., a user searching for a type of geographic entity which is not involved in cluster or co-location rules), because in such cases the additional criteria implemented by *ScoreGR* and *GRBM25* do not influence the output rank, by design. Therefore, in such scenarios, *ScoreGR* and *GRBM25* would resemble the output of the baseline methods, except for the combination of the score of the individual criteria.

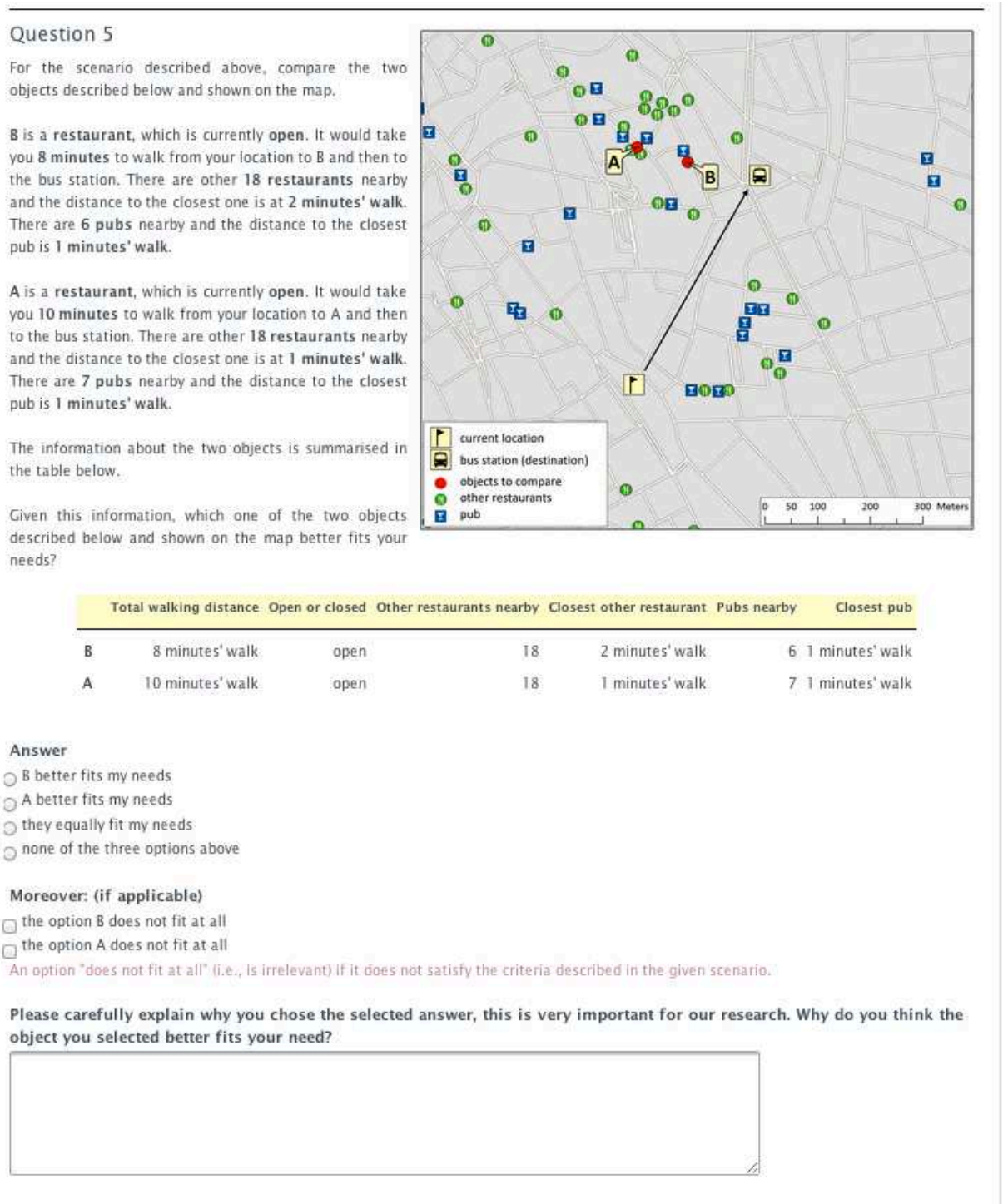
Participants.

A total of 416 participants took part in this experiment in September 2012. Participants were gathered through the online service CrowdFlower (www.crowdflower.com). This service passes the tasks over to the crowdsourcing platform Amazon Mechanical Turk (AMT). We assume that our sample falls into typical AMT demographics (Ross et al., 2010), i.e., computer literate people with no particular expertise in geography. As all participants connected to the service from an U.S. IP address, we considered them to be familiar with

mobile information seeking tasks in an urban context and being valid candidates for our study as they are assumed to represent the general public.

Material.

All three scenarios are set in an urban environment. As map data we used unlabelled Open Street Map data for Madrid, Spain (see Figure 2) including points of interest and a street network (De Sabbata, 2013). We rotated the original map data by 105 degrees counter clockwise and displayed it at a large scale, i.e. approximately 1:8000. With these measures we address the problem that users' familiarity might confound our results since we assume it unlikely that participants recognise the represented geographic area. In the first scenario, a user searches for a supermarket while returning home from work. In the second scenario, a user is searching for a hotel in the area where she is attending a conference. In the third scenario, a user is searching for a restaurant. An example of a question for scenario 3 is shown in Figure 2.



Question 5

For the scenario described above, compare the two objects described below and shown on the map.

B is a **restaurant**, which is currently **open**. It would take you **8 minutes** to walk from your location to B and then to the bus station. There are other **18 restaurants** nearby and the distance to the closest one is at **2 minutes' walk**. There are **6 pubs** nearby and the distance to the closest pub is **1 minutes' walk**.

A is a **restaurant**, which is currently **open**. It would take you **10 minutes** to walk from your location to A and then to the bus station. There are other **18 restaurants** nearby and the distance to the closest one is at **1 minutes' walk**. There are **7 pubs** nearby and the distance to the closest pub is **1 minutes' walk**.

The information about the two objects is summarised in the table below.

Given this information, which one of the two objects described below and shown on the map better fits your needs?

	Total walking distance	Open or closed	Other restaurants nearby	Closest other restaurant	Pubs nearby	Closest pub
B	8 minutes' walk	open	18	2 minutes' walk	6	1 minutes' walk
A	10 minutes' walk	open	18	1 minutes' walk	7	1 minutes' walk

Answer

- ○ B better fits my needs
- ○ A better fits my needs
- ○ they equally fit my needs
- ○ none of the three options above

Moreover: (if applicable)

- ☐ the option B does not fit at all
- ☐ the option A does not fit at all

An option "does not fit at all" (i.e., is irrelevant) if it does not satisfy the criteria described in the given scenario.

Please carefully explain why you chose the selected answer, this is very important for our research. Why do you think the object you selected better fits your need?

Figure 2: Example of judgement task presented to participants on CrowdFlower (De Sabbata, 2013)

Design.

As the number of entities in an Open Street Map dataset is very large, it is not feasible to collect relevance judgements for all entities in a given geographic area. Therefore, we reduced the number of judgements to be made for all three scenarios applying a pooling approach commonly used in IR (Ross et al., 2010). We pre-computed the relevance of all

entities and ranked them for all four methods to be evaluated (see Table 1). In a second step, those entities in the top-k list of at least one of the methods were included in the set of entities to be judged. The underlying assumption is that a relevant geographic entity would be recognised by at least one of the methods. In practice, many of the elements in the top-k lists of the four methods were common to at least two or three methods, a strong indication that relevant entities have been identified by the pooling approach. A manual check of the dataset provided no evidence that any obvious relevant entity had been excluded.

Next, as proposed by (Urbano et al., 2010), we first randomly selected one entity from the unordered set of entities to be judged as a pivot for each scenario (see entity '9698', circled and labelled '1' , in the first line of Figure 3). Inspired by (Manning, Raghavan, & Schütze, 2008) we then created an iteration, that is, a list of randomly ordered, pairwise comparisons between the selected pivot (i.e., entity '9698') and all the remaining entities of the list (see first line of Figure 3). For each comparison, the labels A and B were randomly assigned to two entities to be compared (see Figure 2). Next, in a second iteration the two right-most items of the subgroups were selected as pivots and the remaining entities compared (see entity '9700' circled and labelled '2', and entity '9695' circled and labelled '3', in the second line of Figure 3). Lastly, after a third iteration the list was sorted again according to the relevance of the entities (see entity '9693' circled and labelled '4' and selected as pivot, in the third line of Figure 3). Note that it is not necessary to judge all possible pairs; a subset is enough to get a completely sorted list (see bottom line of Figure 3).

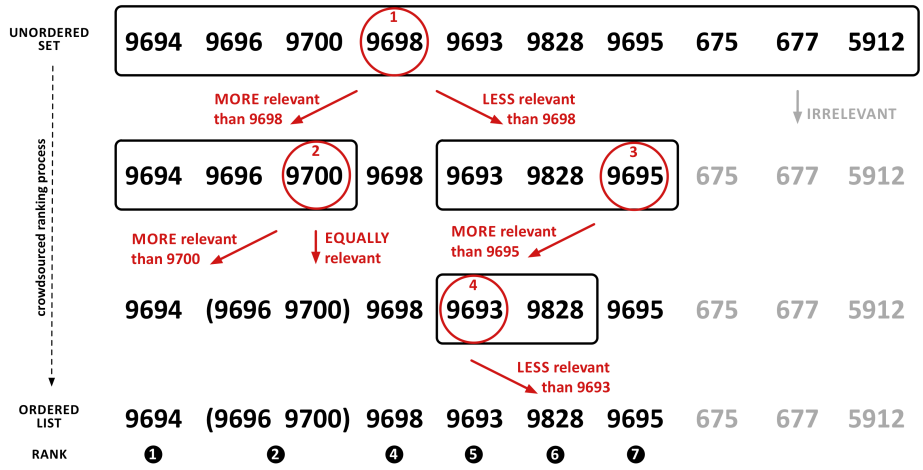


Figure 3: Example of iterations of pair-wise relevance judgements of entities (De Sabbata, 2013) (entities are labelled with their unique identifier, i.e. the bolded 3 or 4 digit codes)

Additionally, each iteration includes up to three check comparisons. Check comparisons are duplicates of one of the comparisons in each iteration, where either the order of presentation of the entities, the labelling of the entity A or B, or both have been swapped. Moreover, we applied a Latin Square design to produce orderings that start with a different comparison, but otherwise follow the same order.

Procedure.

The iterations described above were submitted to CrowdFlower which allocated workers to tasks, i.e., the previously generated iterations (see Section Design earlier). A minimum of 40 participants was allocated per iteration and at least four participants for each distinct order. Figure 2 shows a task example, i.e., one iteration, as presented to a worker on CrowdFlower. In each iteration participants were asked to make a comparative assessment of two entities *A* and *B* (see Figure 2) and select one of the two entities *A* or *B* as more relevant. Participants could also classify entities *A* and *B* as equally relevant, or both as non-relevant. We also requested participants to explain their judgement in a text box. As this entry is mandatory, the collected qualitative data may help to better understand the rationale behind participants' judgements and serves as quality check.

Results and Interpretation

We first transformed the crowdsourced answers into ranks for each scenario, to be used as 'ground truth'. To compare crowdsourced ranks with the ranks generated by *Baseline1*, *Baseline2*, *ScoreGR*, and *GRBM25* we computed the Kendall's τ correlation coefficient. The resulting statistics for the three scenarios are reported in Table 2.

For all three scenarios the correlation between the crowdsourced rank and *ScoreGR* are significant ($p < .05$). For Scenario 2 this is even highly significant ($p < .01$). These results show that crowdsourced relevance assessments (i.e., the assumed ground truth rank) account for 30% of the variability of *ScoreGR* in Scenario 1, 74% in Scenario 2, and 47% in Scenario 3. No significant correlation was found between crowdsourced relevance assessments and any of the baseline methods or *GRBM25* for any scenario.

The results of the experiment show that *ScoreGR* effectively calculates GR for the scenarios we tested. They also reveal that the selected criteria (*topicality*, *spatio-temporal proximity*, *directionality*, *cluster,* and *co-location*) combined to a single GR score can be used to effectively rank geographic entities, as humans would do using the same criteria. Kendall's τ coefficients (Table 2) are only significant for the correlation between the crowdsourced rank and *ScoreGR,* while there are no significant correlations between the crowdsourced rank and *GRBM25,* or the two baselines. This also suggests that latter three models are not adequate for complex scenarios, as tested in this experiment.

For the first two scenarios, *ScoreGR* is able to correctly identify the most relevant geographic entity as judged by users, while for the third scenario it ranks the second most relevant entity first, and vice versa (see Figures 4 and 5, respectively). *ScoreGR* also correctly identifies irrelevant entities in the first and third scenario, i.e., entities that are spatio-temporally not available. Only in Scenario 2 *ScoreGR* yields different results. While study participants classified three geographic entities located very close to the user's position as irrelevant, because they belong to categories not matching the user's need, *ScoreGR* does not

identify these as irrelevant entities. The reason for this is that the underlying measure identifies a semantic similarity between their categories and the user query, although the assigned ranks are very low (i.e., they are classified among the least relevant entities). Whilst *ScoreGR* performs quite well, *Baseline1* and *Baseline2* classify an irrelevant, spatio-temporally not available entity as the top-ranked one, because it is closest to the user's movement path in Scenario 1.

The correlation between *ScoreGR* and *Crowdsource* is lower for Scenario 1 and 3, where participants seem to have weighted the criterion *cluster* slightly higher than *co-location*. However, we treated the criteria *cluster* and *co-location* as equally important, when combining them to compute the geographic environment component of *ScoreGR*. For instance, entity 9115 is ranked second by *ScoreGR*, but fifth by participants in Scenario 1, because it satisfies the co-location rule (pharmacies next to supermarkets) involved in the scenario well, but it is not part of a cluster. In all scenarios, the top-ranked entities belong to a cluster, according to the crowdsourced ranks. A possible explanation for this is, that visual clusters are pre-attentively processed by the human visual system, and thus clusters are very salient (Davies, Fabrikant, & Hegarty, 2015). Aiming for an even better approximation of the human-based ranks, further implementations of the *ScoreGR* method might therefore require a higher importance to be assigned to the criterion *cluster*.

The overall differences between crowdsourced and *ScoreGR* ranks are smaller in Scenario 1 and 2 than in Scenario 3 (see Figure 6). In Scenario 1 we observe that two entities closest to the top ranked show a difference in rank of two and three, respectively (see Figure 6). Differences in ranks are lowest and the spatial distribution of differences more homogeneous in Scenario 2 (see Figure 6). In Scenario 3 we can notice rather large differences in ranks for three entities to the right of the user (see Figure 6), while the remaining six entities show only small differences of one or two ranks.

Scenario 1: searching for a supermarket					
Entity ID	*Crowdsourced*	*Baseline1*	*Baseline2*	*ScoreGR*	*GRBM25*
9128	1	7	7	1	2
9127	2	3	3	4	4
9126	3	5	5	6	7
9124	4	8	8	5	8
9115	5	4	4	2	1
9117	6	2	2	3	3
9125	7	6	6	8	6
9121	8	9	9	7	5
9123	irr	1	1	irr	Irr
		-.111, $p > .05$	-.111, $p > .05$	-.556*, $p < .05$	-.333, $p > .05$
Scenario 2: searching for a hotel					
Entity ID	*Crowdsourced*	*Baseline1*	*Baseline2*	*ScoreGR*	*GRBM25*
9694	1	2	6	1	2
9696	2	5	14	4	5
9700	2	6	16	3	6
9698	4	10	21	2	10
9693	5	3	7	6	3
9828	6	1	2	7	1
9695	7	4	10	8	4
675	irr	Irr	4	206	51
677	irr	Irr	1	193	41
5912	irr	Irr	3	77	40
		-.458, $p > .05$	-.442, $p > .05$	-.861**, $p < .01$	-.442, $p > .05$
Scenario 3: searching for a restaurant					
Entity ID	*Crowdsourced*	*Baseline1*	*Baseline2*	*ScoreGR*	*GRBM25*
714	1	1	1	2	1
704	2	5	5	1	3
7212	3	13	13	5	13
7213	3	12	12	4	9
724	5	3	3	38	4
7211	5	19	19	3	20
747	7	2	2	15	2
746	8	7	7	17	5
711	irr	4	4	Irr	Irr
		-.057, $p > .05$	-.057, $p > .05$	-.686*, $p < .05$	-.400, $p > .05$

Table 2: Comparison between crowdsourced and computed ranks for the three scenarios. Note: the label 'irr' refers to entities identified as irrelevant; * indicates significance on 5% and ** on 1 % level.

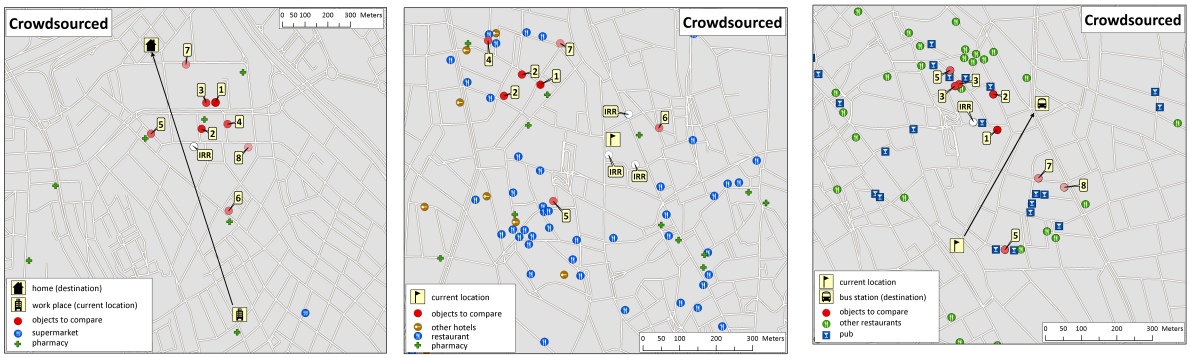


Figure 4: Ranking of the pooled entities from *Crowdsource* judgements for Scenario 1, 2, and 3.
Note: The red coloured circles represent the ranked entities; 'IRR refers to entities identified as irrelevant. In Scenario 1 (left) and Scenario 3 (right) the arrow indicates the user's route direction from start to end; the flag represents the user's current position in Scenario 2 (middle) and Scenario 3 (right). In Scenario 1 (left) the blue circles represent other supermarkets, and the green crosses represent pharmacies. In Scenario 2 (middle) the brown circles represent other hotels, the blue circles restaurants, and the green crosses pharmacies. In Scenario 3 (right) the green circles represent other restaurants and the blue symbols represent pubs.

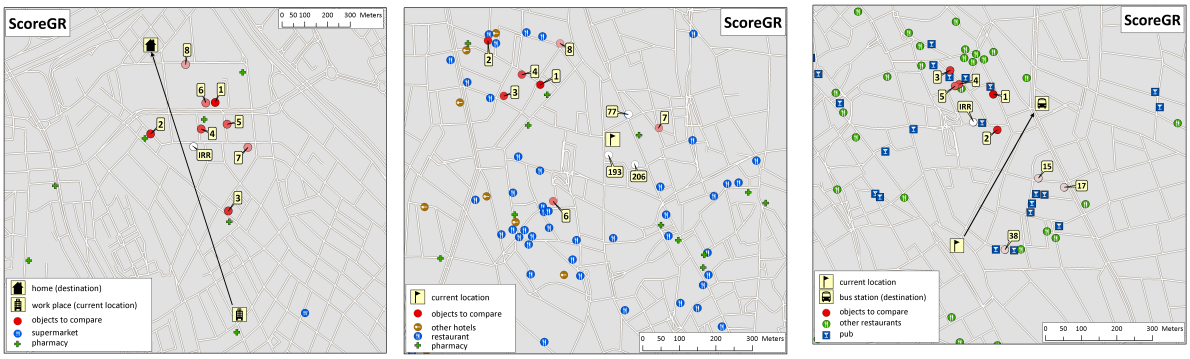


Figure 5: Ranking of the pooled entities with *ScoreGR* method for Scenario 1, 2, and 3 (see notes Fig. 4)

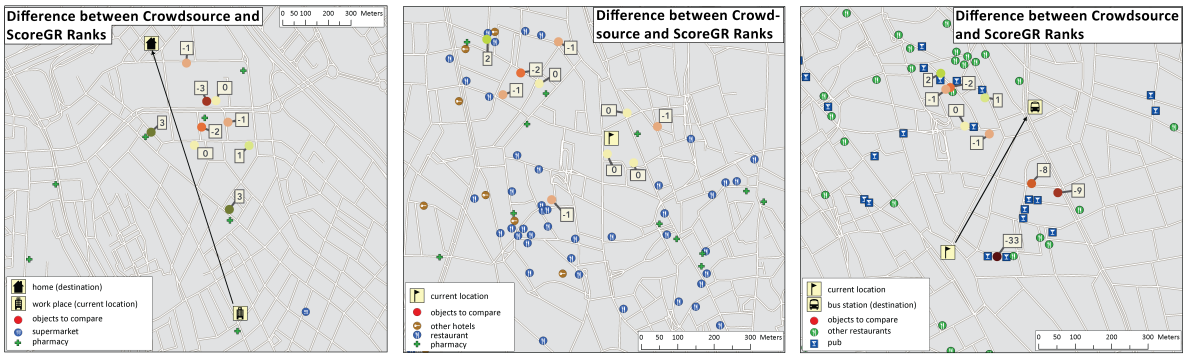


Figure 6: Differences of ranks between *Crowdsource* and *ScoreGR* for Scenario 1, 2 and 3.
Note: The coloured circles represent the differences in ranks between crowdsourced judgements (Fig. 4) and ranks obtained with the ScoreGR method (Fig. 5). Green colours indicate positive, red colours negative deviations. Yellow circles stand for no difference.

While the correlation coefficients (see Table 2) clearly show that *ScoreGR* and the crowdsourced ranks correlate, and thus *ScoreGR* effectively assesses GR in the three scenarios, this is not the case for *GRBM25*. This method assumes that the probability of satisfying a user's information need is dependent on the number of entities that are closer to, or at the same distance from the user's information need for each criterion (for details about

mapping criteria to distances and calculating the distance values, see De Sabbata, 2013). This assumption though yields an undesired dominance of criteria with higher variability over those criteria that tend to produce tied values, such as *clusters*. Therefore, the probability scores computed for the criteria *topicality*, *spatio-temporal proximity*, and *directionality* have a dominant influence on the final score produced by *GRBM25* with respect to the probability scores computed for the criteria *cluster* and *co-location*. This was confirmed by computing an additional baseline method, which filters the geographic entities by category and ranks them based on their *spatio-temporal proximity* score. In fact, there is no significant correlation between the crowdsourced ranks and this additional third baseline, but the latter shows a highly significant correlation ($p < .01$) with *GRBM25* for all the three scenarios (with Kendall's correlation coefficients $\tau > .90$ for Scenarios 2 and 3). This result also confirms that the criteria *cluster* and *co-location* have to be treated separately, as they cannot be captured using more simple spatial criteria, such as spatial proximity.

Discussion

Overall, our crowdsourcing results show that our proposed *ScoreGR* method is capable of effectively assessing GR for the implemented scenarios. Human relevance ranking is similar to rankings computed by *ScoreGR*. The correlation ranged from .50 to .851 for the three scenarios. For scenarios 1 and 3, where participants weighted the criterion *co-location* lower than *cluster* the strength of correlation is weaker. As *ScoreGR* was designed to weight these two criteria equally, the observed difference suggests that *cluster* plays a more important role than *co-location* when humans assess GR. At the same time it also implies that the importance of these two criteria may vary for different scenarios. Overall, the results suggest that *cluster* and *co-location* are essential when computing the final GR score and that the importance we assigned to them in the combination method of *ScoreGR* is too low compared to the crowdsourced ranks. Entities that highly satisfy the *cluster* and *co-location* criteria received higher crowdsourced ranks than the computed *ScoreGR* ranks. At the same

time, both baseline methods produced no similar ranks with respect to human judgements. They perform better in Scenario 2, where *spatio-temporal proximity* is reduced to spatial distance (as the time factor is not important in that scenario), and when *directionality* is excluded. However, even in Scenario 2 the ranks produced by the baselines methods do not resemble the crowdsourced ranks. These empirical results are coherent with previous work and conceptualisations of GR (De Sabbata & Reichenbacher, 2012; Raper, 2007; Reichenbacher, 2007; Zipf, 2003).

Although commonly used criteria such as in IR (*topicality*) (e.g., Greisdorf, 2000; Sanderson, 2010; Schamber, Eisenberg, & Nilan, 1990), in LBS (*spatial proximity*) (Brimicombe, 2008), or in GIR (*topicality* and *spatial proximity*) (Mandl, 2011) seem appropriate and effective by respective applications, crowdsourced responses from our experiment indicate that they are not sufficient for understanding GR. Spatial proximity (represented by *Baseline1* and *Baseline2*) commonly implemented in current mobile information services, such as LBS, show weak correlations with human judgements, and suggest that spatial distance alone is not enough for assessing the relevance of geographic entities for specific context-dependent and realistic, information needs.

Crucially, most mobile information needs implicitly entail a temporal dimension. For instance, if a user intends to go grocery shopping and asks ‘show me the closest supermarket’, they implicitly assume that a supermarket should not only be nearby, but also open. If an entity is not spatio-temporally accessible with respect to a user’s space-time constraints, the entity is not relevant for the shopping activity. In such cases, participants ranked the obsolete entities as irrelevant. This implies that spatial relevance needs must to be considered with temporal relevance, i.e., the time a user intends to or needs to spend for a particular activity at a specific location.

The results of our experiment not only confirm our previous studies (De Sabbata & Reichenbacher, 2012), they also support the theoretical relevance concepts proposed by

(Raper, 2007; Reichenbacher, 2007), and underline the main claim by Raper (2007, p. 837) that ‘situational relevance concepts as currently articulated do not deal sufficiently with concepts of mobility and geography, and that these concepts are essential to the understanding of mobile information seeking’. At the same time our empirical results do not support another claim made by Raper (2007, p. 842–843) that whatever is in the vicinity of the user is topically relevant, simply because it is in the ‘accessibility envelope or surroundings of the user’. The relevance judgements collected in our experiment clearly show that geographic entities close to a user’s current position are assessed as non-relevant, if their semantics are not relevant to the user task at hand.

One reason for the poor performance of *Baseline2* is that such methods of weighted linear combinations or geometric combinations of relevance scores are too rigid to account for the different natures of the relevance criteria. They tend to overestimate the final score by weighing the spatial score too strongly and by lacking compensation. In Scenario 2, for instance, *Baseline2* ranks entities that are close to the user as very relevant, even though they do not match the semantic category the user is looking for. This underlines the general difficulty of combining topical and spatial scores. *Baseline1* is obviously not affected by this particular problem, since the rigidity of filter-based methods classifies semantically dissimilar entities as irrelevant in the first place (e.g., motels and hostels would be considered as completely irrelevant, when searching for a hotel). This binary, disjunctive combination is non-compensatory, and does not reflect human judgements. Our findings support the superiority of more sophisticated methods, such as the Continuous Preference Logic model (Dujmovic, 2007) used in *ScoreGR* that allow for fuzzy conjunctive combinations and conjunctive partial absorptions. These methods are less rigid than filter-based approaches and allow for compensation.

Our experiment also highlights the importance of an adequate approximation of *topicality*. While simple category filtering is computationally efficient and often an adequate

approach, by relying on syntax only, it can be too exclusive and discard too many entities as non-relevant. We recommend using semantic similarity to *topicality* instead of an exclusive filtering based on category labels (e.g., Miller & Charles, 2007). For example, instead of excluding all entities that are not of the category 'hotel', entities of similar kinds, e.g., 'hostel', 'guesthouse', 'motel' should be treated as semantically similar and be included as partly relevant. It is fair to assume that most humans would accept certain trade-offs in terms of space, semantics, and functionality, e.g., a nearby motel versus a distant hotel. This is supported by various empirical studies that show that similarity is context-dependent and malleable (e.g., Goldstone, 1994). However, one key problem remains as how to define and implement semantic similarities of geographic features. The solution implemented for *ScoreGR* is based on the Normalized Google Distance (Cilibrasi & Vitanyi, 2007), as detailed in (De Sabbata, 2013), and offers plausible and meaningful scores for *topicality*, but is strongly dependent on the underlying dataset. It can also produce inadequate results, for example that pubs are almost as similar to hotels as hostels are to hotels. Topical relevance assessment becomes even more difficult, if the user cannot explicitly specify a category name (e.g., 'restaurant'), but only her intention or objective for a planned activity (e.g., 'dinner'). In general, a possible solution to such problems can be to actually assume the objective behind a user query, deduce possible activities, and match them with respective affordances of the geographic features. However, such a solution requires sophisticated activity analysis and a systematic understanding of common-sense knowledge (Gunning, Chaudhri, & Welty, 2010; Liu & Singh, 2004; Singh et al., 2002), especially concerning affordances associated with different types of geographic features (Alazzawi, Abdelmoty, & Jones, 2012; Alves & Pereira, 2012).

An accurate assessment of *spatio-temporal proximity* is crucial to capture *topicality* well. It has been shown by (Boscoe, Henry, & Zdeb, 2012) that direct distance is a good first estimation of distances along a route network. Although we could show that direct distance

implemented in *ScoreGR* is an effective and efficient approximation for spatio-temporal accessibility in our experiment, it may be too simple in other cases, especially where movement is constrained by a network. It is limited by the assumption of uniform speed of movement that hardly holds in reality. For mobile users, the relevance assessment should at least discern different means and modes of transport, including public transport, and real-time traffic situations. As any spatio-temporal accessibility assessment method is an estimation, and includes computational costs for increased accuracy. Further empirical studies will have to consider cost/benefit ratios for degrees of increased accuracy and associated computational costs.

Our findings support both the assumed static influence of the spatial layout of the geographic features and the dynamic influence of the user's context and mobility on GR assessment. This is coherent with prior conceptualisations of static and dynamic context in the literature (e.g., Dey, 2001; Hong, Suh, & Kim, 2009; Kofod-Petersen & Cassens, 2006; Schmidt et al., 1999). The static GR component corresponds to the *geographic environment* in our computational model (see Fig. 1). As part of the context component of relevance (Coppola et al., 2004; Mizzaro, 1998) it encompasses the spatial configuration and spatial relations of entities beyond spatial proximity, such as co-locations, clusters, connectedness, etc. Our results also confirm that the relevance of a single entity increases, if there are several entities of the same category in a neighbourhood (*cluster*) (De Sabbata & Reichenbacher, 2012) or if it is located next to an entity of a related category that is typically in close spatial proximity (*co-location*) (De Sabbata & Reichenbacher, 2012). These criteria are 'the differences in situational contexts and research task requirements' (Barry & Schamber, 1998, p. 234) that separate GR from the concept of relevance commonly applied in classic document-based IR. We believe that the geographic environment, i.e., the spatial configuration of geographic features defines a kind of basic or elementary geographic relevance.

This stable geographic relevance is strongly influenced and modified by a second, dynamic component of GR. The dynamics of the environment and the mobility of a user can further modulate the basic, configurational geographic relevance through direction of movement or spatio-temporal accessibility of entities. This component corresponds to *mobility* in our computational model (see Fig. 1). We understand mobility as a means to perform activities at various places, thus spatio-temporal proximity of entities is central to assessing GR with respect to such mobile user activities. Results for Scenario 1 indicate that planning may also play a substantial role in GR assessment. A few participants judged a supermarket further away from the destination and with less time for shopping as more relevant than one, which was open longer and located closer to the destination. Comments by these participants reveal that they planned to go to the second – and still open – supermarket on their way to the destination, if they could not find what they were looking for in the first one. Although it is hard to generalise, it suggests that for modelling GR, the geographic environment of entities is as important as the context in which the user is seeking information to satisfy her needs.

Although our results suggest the validity of the applied criteria of geographic relevance, our experimental design only used simulated work tasks in the sense of (Borlund & Schneider, 2010). While these tasks are certainly plausible and prototypical for everyday information seeking tasks, future studies will have to carefully design tasks that capture information needs in the real world and employ tasks that participants will accomplish in the real world. Selection of such tasks could be informed by an analysis of mobile search queries. In particular, such tasks should cover more complex mobile situations and information needs triggered by linked activities and influenced by their dynamic coordination and planning. However, this would require field studies with even less control and fewer participants.

Conclusions

Integrating the findings from our previous work (De Sabbata & Reichenbacher, 2012) with the results obtained from the experiment presented here, we argue that geographic relevance expresses multi-faceted relationships between a mobile user's geographic information needs and the geographic entities in the user's environment. The two main criteria defining the strength of the GR relationship are the spatio-temporal accessibility of an entity with respect to a user's mobility (*spatio-temporal proximity*), and the topicality of an entity's affordances with respect to the information need for a particular activity. Furthermore, the strength of the GR relationship is strongly influenced by the geographic context of an entity's location, such as spatial clusters and co-location of other relevant geographic entities.

Therefore, GR is distinct from the concept of relevance commonly used in IR. Our empirical data indicate that GR cannot be adequately calculated by simply combining category filtering and direct distance-based ranking, i.e., *Baseline1 and Baseline2*. In contrast, our proposed method *ScoreGR*, proved to be effective in assessing GR for the considered scenarios.

Furthermore, we found that crowdsourcing was a useful complementary approach for testing GR assessment methods, beyond controlled lab studies. Although the outcome of the experiment overall supports our claim that GR is distinct from concepts of relevance in IR and GIR, and that *ScoreGR* is an effective and valid method for assessing GR, we need further and extended empirical studies. Extensions should not only encompass further criteria and different scenarios, but additional propositions on combining scores.

Future studies could focus on evaluating GR methods in scenarios where the criteria *cluster* or *co-location* are not as influential for the GR score, in order to test for the robustness of the proposed methods. For instance, the discussed methods can be tested in a scenario depicting a user searching for a hotel, in an area where hotels are not clustered and not

satisfying any of the co-location rules. Finally, we recommend testing the validity of the GR methods described as well as the criteria they are based on, in field studies with users moving in the real world.

Acknowledgments

The presented work is part of the project ‘Geographic Relevance in Mobile Applications’ funded by the Swiss National Science Foundation (Project 200021_119819 / 1).

We would like to thank Omar Alonso from Microsoft Inc., for his support with using Amazon Mechanical Turk.

References

Alazzawi, A. N., Abdelmoty, A. I., & Jones, C. B. (2012). What can I do there? Towards the automatic discovery of place-related services and activities. *International Journal of Geographical Information Science, 26*(2), 345-364.

Alonso, O., & Baeza-Yates, R. (2011). Design and implementation of relevance assessments using crowdsourcing. In P. Clough, C. Foley, C. Gurrin, G. F. Jones, W. Kraaij, H. Lee & V. Mudoch (Eds.), *Advances in Information Retrieval, Lecture Notes in Computer Science Volume 6611* (pp. 153-164). Berlin, Heidelberg: Springer.

Alonso, O., & Mizzaro, S. (2009). Can we get rid of TREC assessors? Using Mechanical Turk for relevance assessment, *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation* (pp. 15-16).

Alonso, O., Rose, D. E., & Stewart, B. (2008). Crowdsourcing for relevance evaluation. *ACM SIGIR Forum, 42*(2), 9-15.

Alves, A. O., & Pereira, F. C. (2012). Making sense of location context, *Proceedings of the 1st International Workshop on Context Discovery and Data Mining* (pp. 4).

Andrade, L., & Silva, M. J. (2006). Relevance ranking for geographic IR, *3rd Workshop on Geographic Information Retrieval at SIGIR'06*. Seattle, Washington, USA.

Barry, C. L., & Schamber, L. (1998). Users' criteria for relevance evaluation: a cross-situational comparison. *Information Processing & Management, 34*(2-3), 219-236.

Borlund, P., & Schneider, J. W. (2010). Reconsideration of the simulated work task situation: A context instrument for evaluation of information retrieval interaction. In N. J. Belkin & D. Kelly (Eds.), *Proceeding of the Third Symposium on Information Interaction in Context (IIiX'10)* (pp. 155-164): ACM: New York.

Boscoe, F. P., Henry, K. A., & Zdeb, M. S. (2012). A nationwide comparison of driving distance versus straight-line distance to hospitals. *The Professional Geographer, 64*(2), 188-196.

Brimicombe, A. J. (2008). Location-Based Services and Geographic Information Systems. In J. P. Wilson & A. S. Fotheringham (Eds.), *The Handbook of Geographical Information Science* (pp. 581-595). Oxford: Wiley-Blackwell.

Cai, G. (2011). Relevance ranking in Geographical Information Retrieval. *The SIGSPATIAL Special, 3*(2), 33-36.

Carvalho, V. R., Lease, M., & Yilmaz, E. (2011). Crowdsourcing for search evaluation. *SIGIR Forum, 44*(2), 17-22.

Cilibrasi, R. L., & Vitanyi, P. M. B. (2007). The google similarity distance. *Knowledge and Data Engineering, IEEE Transactions on, 19*(3), 370-383.

Clough, P., Joho, H., & Purves, R. (2006). Judging the Spatial Relevance of Documents for GIR. In M. Lalmas, A. MacFarlane, S. Rüger, A. Tombros, T. Tsikrika & A. Yavlinsky (Eds.), *Advances in Information Retrieval, Lecture Notes in Computer Science Volume 3936* (pp. 548-552). Berlin, Heidelberg: Springer.

Clough, P., Sanderson, M., Jiayu, T., Gollins, T., & Warner, A. (2013). Examining the limits of crowdsourcing for relevance assessment. *IEEE Internet Computing, 17*(4), 32-38.

Coppola, P., Della Mea, V., Di Gaspero, L., & Mizzaro, S. (2004). The concept of relevance in mobile and ubiquitous information access. In F. Crestani, M. Dunlop & S. Mizzaro (Eds.), *Mobile and Ubiquitous Information Access, Lecture Notes in Computer Science Volume 2954* (pp. 1-10). Berlin, Heidelberg: Springer.

Cosijn, E., & Ingwersen, P. (2000). Dimensions of relevance. *Information Processing & Management, 36*(4), 533-550.

Crowley, J., Coutaz, J., Rey, G., & Reignier, P. (2002). Perceptual components for context aware computing. In G. Borriello & L. Holmquist (Eds.), *UbiComp 2002: Ubiquitous Computing, Lecture Notes in Computer Science Volume 2498* (pp. 117-134). Berlin, Heidelberg: Springer.

Davies, C., Fabrikant, S. I., & Hegarty, M. (2015). Towards Empirically Verified Cartographic Displays. In J. Szalma, M. Scerbo, P. Hancock, R. Parasuraman & R. Hoffman (Eds.), Cambridge Handbook of Applied Perception Research (pp. 711-729). New York, NY: Cambridge University Press.

De Sabbata, S. & Reichenbacher, T. (2014). Computing geographic relevance in mobile information services. *Proceedings of the 1st International Workshop on Context-Awareness in Geographic Information Services (CAGIS 2014)* in conjunction with GIScience 2014 (CAGIS 2014), 17-28.

De Sabbata, S. (2013). *Assessing Geographic Relevance for Mobile Information Services.* Unpublished Doctoral Thesis, University of Zurich, Zurich.

De Sabbata, S., & Reichenbacher, T. (2012). Criteria of geographic relevance: an experimental study. *International Journal of Geographical Information Science, 26*(8), 1495-1520.

Dey, A. K. (2001). Understanding and using context. *Personal and Ubiquitous Computing, 5*(1), 4-7.

Dujmovic, J. (1975). Extended continuous logic and the theory of complex criteria. *Journal of the University of Belgrade, Series on Mathematics and Physics, 537*, 197-216.

Dujmovic, J. (2007). Continuous preference logic for system evaluation. *IEEE Transactions on Fuzzy Systems, 15*(6), 1082-1099.

Dujmovic, J., & Larsen, H. (2007). Generalized conjunction/disjunction. *International Journal of Approximate Reasoning, 46*(3), 423-446.

Eickhoff, C., & de Vries, A. (2011). How crowdsourcable is your task, *Proceedings of the Workshop on Crowdsourcing for Search and Data Mining (CSDM) at the Fourth ACM International Conference on Web Search and Data Mining (WSDM)* (pp. 11-14). Hong Kong, China.

Felstiner, A. (2011). WorkingThe Crowd: Employment And Labor Law In The Crowdsourcing Industry. *Berkeley Journal of Employment and Labor Law, 32*(1), 143-204.

Goldstone, R. (1994). The role of similarity in categorization: providing a groundwork. *Cognition, 52*(2), 125-157.

Greisdorf, H. (2000). Relevance: An interdisciplinary and information science perspective. *Informing Science, 3*(2), 67-72.

Gunning, D., Chaudhri, V. K., & Welty, C. (2010). Introduction to the Special Issue on Question Answering. *AI Magazine, 31*(3), 11-12.

Harris, C. (2011). Youre Hired! An Examination of Crowdsourcing Incentive Models in Human Resource Tasks, *Proceedings of the Workshop on Crowdsourcing for Search and Data Mining (CSDM) at the Fourth ACM International Conference on Web Search and Data Mining (WSDM)* (pp. 15-18). Hong Kong, China.

Hong, J., Suh, E., & Kim, S. J. (2009). Context-aware systems: A literature review and classification. *Expert Systems with Applications, 36*(4), 8509-8522.

Howe, J. (2006). The Rise of Crowdsourcing. *Wired, Issue 14.06 - June 2006*.

Huang, H., & Gartner, G. (2009). Using activity theory to identify relevant context parameters. In G. Gartner & K. Rehrl (Eds.), *Location Based Services and TeleCartography II* (pp. 35-45). Berlin, Heidelberg: Springer.

Huang, Y., Shekhar, S., & Xiong, H. (2004). Discovering colocation patterns from spatial data sets: a general approach. *Knowledge and Data Engineering, IEEE Transactions on, 16*(12), 1472-1485.

Ipeirotis, P. (2010a). Analyzing the Amazon Mechanical Turk Marketplace. *XRDS, 17*(2), 16-21.

Ipeirotis, P. (2010b). Demographics of Mechanical Turk, *Center for Digital Economy Research, Working paper No. CEDER-10-01* New York University, Stern School of Business.

Jiang, B., & Yao, X. (2006). Location-based services and GIS in perspective. *Computers, Environment and Urban Systems, 30*(6), 712-725.

Jones, C. B., & Purves, R. S. (2008). Geographical information retrieval. *International Journal of Geographical Information Science, 22*(3), 21-28.

Kofod-Petersen, A., & Cassens, J. (2006). Using activity theory to model context awareness. In T. Roth-Berghofer, S. Schulz & D. Leake (Eds.), *Modeling and Retrieval of Context, Lecture Notes in Computer Science Volume 3946* (pp. 1-17). Berlin, Heidelberg: Springer.

Kumar, C. (2011). Relevance and ranking in geographic information retrieval, *Proceedings of FDIA'11, Fourth BCS-IRSG conference on Future Directions in Information Access* (pp. 2-7). Koblenz, Germany: British Computer Society.

Liu, H., & Singh, P. (2004). ConceptNet – a practical commonsense reasoning tool-kit. *BT Technology Journal, 22*(4), 211-226.

Mandl, T. (2011). Evaluating GIR: geography-oriented or user-oriented? *The SIGSPATIAL Special, 3*(2), 42-45.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge: Cambridge University Press.

Marsden, P. (2009). Crowdsourcing. *Contagious Magazine, 18*, 24-28.

Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods, 44*(1), 1-23.

McCreadie, R. M. C., Macdonald, C., & Ounis, I. (2010). Crowdsourcing a news query classification dataset. In M. Leas, V. Carvalh & E. Yilmaz (Eds.), *Proceedings of the*

ACM SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation (CSE 2010) (pp. 31-38). Geneva, Switzerland: ACM.

Miller, G. A., & Charles, W. G. (2007). Contextual correlates of semantic similarity. *Language and Cognitive Processes, 6*(1), 1-28.

Miller, H. J., & Bridwell, S. A. (2009). A field-based theory for time geography. *Annals of the Association of American Geographers, 99*(1), 49-75.

Mizzaro, S. (1998). How many relevances in information retrieval? *Interacting With Computers, 10*(3), 303-320.

Mountain, D., & Macfarlane, A. (2007). Geographic information retrieval in a mobile environment: evaluating the needs of mobile individuals. *Journal of Information Science, 33*(5), 515-530.

Purves, R., Clough, P., Jones, C. B., Arampatzis, A., Bucher, B., Finch, D. … Yang, B. (2007). The design and implementation of SPIRIT: a spatially aware search engine for information retrieval on the Internet. *International Journal of Geographical Information Science*, 21(7), 717-745.

Raper, J. (2007). Geographic relevance. *Journal of Documentation, 63*(6), 836-852.

Raper, J., Gartner, G., Karimi, H., & Rizos, C. (2007a). Applications of location-based services: a selected review. *Journal of Location Based Services, 1*(2), 89-111.

Raper, J., Gartner, G., Karimi, H., & Rizos, C. (2007b). A critical evaluation of location based services and their potential. *Journal of Location Based Services, 1*(1), 5-45.

Raubal, M., Miller, H. J., & Bridwell, S. (2004). User-centred time geography for location-based services. *Geografiska Annaler Series B Human Geography, 86*(4), 245-265.

Reichenbacher, T. (2007). The concept of relevance in mobile maps. In G. Gartner, W. Cartwright, M. P. Peterson, W. Cartwright, G. Gartner, L. Meng & M. P. Peterson (Eds.), *Location Based Services and TeleCartography* (pp. 231-246). Berlin, Heidelberg: Springer.

Ross, J., Irani, L., Silberman, M. S., Zaldivar, A., & Tomlinson, B. (2010). Who are the crowdworkers?: shifting demographics in Mechanical Turk, *CHI '10 Extended Abstracts on Human Factors in Computing Systems* (pp. 2863-2872). New York: ACM.

Sanderson, M. (2010). Test Collection Based Evaluation of Information Retrieval Systems. *Information Retrieval, 4*(4), 247-375.

Saracevic, T. (1996). Relevance reconsidered, *Proceedings of the 2nd Conference on Conceptions of Library and Information Science* (pp. 201-218): Royal School of Librarianship.

Saracevic, T. (2007). Relevance: a review of the literature and a framework for thinking on the notion in information science. Part II: nature and manifestations of relevance. *Journal of the American Society for Information Science and Technology, 58*(13), 1915-1933.

Schamber, L., Eisenberg, M. B., & Nilan, M. S. (1990). A re-examination of relevance: toward a dynamic, situational definition. *Information Processing & Management, 26*(6), 755-776.

Schmidt, A., Beigl, M., & Gellersen, H. W. (1999). There is more to context than location. *Computers & Graphics, 23*(6), 893-901.

Singh, P., Lin, T., Mueller, E., Lim, G., Perkins, T., & Li Zhu, W. (2002). Open Mind Common Sense: Knowledge acquisition from the general public. In R. Meersman & Z. Tari (Eds.), *On the Move to Meaningful Internet Systems 2002: CoopIS, DOA, and ODBASE, Lecture Notes in Computer Science Volume 2519* (pp. 1223-1237). Berlin, Heidelberg: Springer.

Soleymani, M., & Larson, M. (2010). Crowdsourcing for affective annotation of video: Development of a viewer-reported boredom corpus. In M. Leas, V. Carvalh & E.

Yilmaz (Eds.), *Proceedings of the ACM SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation (CSE 2010)* (pp. 4-8). Geneva, Switzerland: ACM.

Spärck Jones, K., Walker, S., & Robertson, S. E. (2000). A probabilistic model of information retrieval: development and comparative experiments:: Part 2. *Information Processing & Management, 36*(6), 809-840.

Urbano, J., Morato, J., Marrero, M., & Martín, D. (2010). Crowdsourcing preference judgments for evaluation of music similarity tasks. In M. Leas, V. Carvalh & E. Yilmaz (Eds.), *Proceedings of the ACM SIGIR 2010 Workshop on Crowdsourcing for Search Evaluation (CSE 2010)* (pp. 9-16). Geneva, Switzerland: ACM.

Van Kreveld, M., Reinbacher, I., Arampatzis, A., & Van Zwol, R. (2005). Multi-Dimensional Scattered Ranking Methods for Geographic Information Retrieval. *Geoinformatica, 9*(1), 61-84.

Wallgrün, J. O., Yang, J., & Klippel, A. (2014). Cognitive Evaluation of Spatial Formalisms: Intuitive Granularities of Overlap Relations. *International Journal of Cognitive Informatics and Natural Intelligence, 8*(1), 1-17.

Wilson, P. (1973). Situational relevance. *Information Storage and Retrieval, 9*(8), 457-471.

Yuen, M. C., King, I., & Leung, K. S. (2011). A survey of crowdsourcing systems, *Proceedings of Privacy, Security, Risk and Trust (PASSAT), IEEE Third International Confernece on Social Computing (SocialCom)* (pp. 766-773). Boston, MA.

Zipf, A. (2003). Die Relevanz von Geoobjekten in Fokuskarten. In J. Strobl, T. Blaschke & G. Griesebner (Eds.), *Angewandte Geographische Informationsverarbeitung XV: Beiträge zum AGIT-Symposium Salzburg 2003* (pp. 567-576). Heidelberg: Wichmann.